WORKING PAPER

The Gradient and the Source

How a language model inherits the human record through a chain of conditioned projections, why the source it points at is never in the record, and what its use is doing to us

A.H.

2026

Abstract. A large language model is built from the human textual record and is commonly described as a compressed but representative store of human knowledge and experience. The relationship is better understood as a cascade of lossy projections. Experience is projected into expression by a conditioned mind writing toward an effect on a reader who is largely invented; expression is selected into the durable record under conditions that admit only a narrow band of it; the record is re-encoded into a model's numerical substrate by a further conditioned process; and the model, now a conditioned interpreter in its own right, projects again into text that a human reader interprets through one more conditioning. At every transfer there is loss, and a bias toward the local direction of the act rather than toward the source it descended from. Two of these losses must be kept apart. One is a severe selection bias: the record is a skewed sample of human output, dense with what people externalize and nearly empty of what they do not, and it is not relieved by scale. The other is a difference in kind: a record of what people did is not the interior they did it from, the deepest of that interior never becomes language at all, and a system that has only the form of the record, with no grounding in the things the words refer to, cannot recover the source as a grounded human reader can. The paper traces this cascade, then the dynamics that make it self-reinforcing: optimization for fluency, which can be measured, over truth, which cannot; a drift toward the high-probability center of language and away from the rare tail where new things live; and a loop in which output floods the record, re-enters training, and reshapes the people who read it. The danger argued for is not that the machine has no interior but that its use is narrowing the interior of the people who are its only source, for which evidence has begun to accumulate. The source was never in the record, which is why the machine cannot corrupt it, and why the only thing that can lose it is the cessation of the human activity that produces it.

1. Introduction

A person can now put a question to a machine and receive, in seconds, a fluent and often correct answer on almost any subject, in almost any voice, shaped with precision to what was asked. The capability is real, it is useful, and it is not a trick. It is also being adopted faster than any communication technology before it. Google's summarizing system, which answers above the search results before a user clicks, reached more than two billion people a month by the middle of 2025. The most used conversational system is consulted by roughly nine hundred million people a week and handles on the order of two and a half billion prompts a day. In under three years a single class of system has become one of the principal ways the species puts language in front of itself, and the question of what passes through that channel, and what cannot, has stopped being academic.

These systems are built from the human textual record, and the usual description of that record, as a compressed but representative sample of human knowledge and experience, is the thing this account means to correct. The record is not a sample of experience. It is the last visible stage of a chain of projections, each one lossy, each one bent by a conditioned interpreter standing in its middle. Experience becomes expression only by passing through a mind whose perception and articulation are themselves shaped by everything that mind has undergone. Expression becomes part of the durable record only by surviving a set of conditions that admit a narrow band of it. The record becomes a model only by being re-encoded into weights and vectors, a translation performed by a training process and by the human judgments that steer it. The model becomes speech only by projecting once more, and that speech becomes meaning only by being read by a person who interprets it through a conditioning of their own.

A single image holds the whole of it. A gradient points in the direction of steepest local change. It is exact about the slope underfoot and silent about where the summit lies. Each step in this chain is a gradient of that kind: faithful to the local pressures of the act, to the effect being sought and the conditioning of the interpreter, and uninformative about the source the act descended from. The model sits at the end of the chain, a gradient of a gradient of a gradient, and a gradient does not point back at its source.

Two of the losses in this chain have to be separated, because running them together produces both false alarm and false comfort. The first is a severe selection bias. The record is a skewed sample of the things people put into the world, and no amount of further collection makes the sample representative, because the skew is built into the act of recording rather than into how much of it was gathered. The second is a

difference in kind. Even a complete record of what people did would not be the interior they did it from; the deepest part of that interior never takes the form of language at all; and a system that possesses only the form of the record, with no grounding in the things the words refer to, stands in a different relation to meaning than a human reader does. The first loss is, in principle, contingent, and in practice unrepairable. The second does not move at all.

One question is set aside throughout, deliberately and from the start: whether there is anything it is like to be one of these systems, whether anyone is home in the machine. That is the unsolved problem of consciousness, and the argument is built so that it holds whichever way the question falls. What is at issue is not the machine's interior but the human one, which was never in the record to begin with. Where a claim is secure it is stated plainly; where it is contested or interpretive it is marked as such. On a subject this distorted by confidence in both directions, the marking is most of the honesty.

2. What the machine is made of

The construction, at the level that matters here, is simple. A model of this kind is trained by exposure to a corpus of human text larger than any person could read in many lifetimes, until it can extend any passage with the most probable continuation, and then the next, and the next. Everything it later appears to know is a residue of statistical regularities drawn from that corpus. Before it is anything else it is a statistical model of the human written record, and its internal substance is quantitative: weights, and the high dimensional vectors in which it represents words and the relations among them.

The founding theory of communication set the relevant dimension aside at the outset, on purpose, and to great effect. Defining the problem of transmitting a message in 1948, Claude Shannon wrote that *these semantic aspects of communication are irrelevant to the engineering problem*; what mattered was only that a message is one selection from a set of possible ones. The judgment was correct, and the digital world rests on it. The shape of the move is worth seeing. From the beginning, the mathematics that lets us store, transmit, and now generate language at scale was a mathematics in which meaning had been placed, deliberately, outside the problem. The systems inherited that boundary at the root. They operate with great power on the selection and have no hold on the thing the selection was made to convey.

The chain that this paper traces can be set down plainly, because each later section is the examination of one of its links. Experience becomes expression, by way of a conditioned mind that is also, in writing, addressing someone. Expression becomes record, by way of selection. Record becomes model, by way of a numerical re-encoding steered by conditioned human judgment. Model becomes output, by way of a system that is now itself a conditioned interpreter. Output becomes meaning, by way of a reader who supplies a great deal from a life of their own. Every arrow is a projection, every projection loses, and at the origin of every projection stands an interpreter whose conditioning decides what crosses and in what shape. The losses are of two kinds, and they are taken in turn: the contingent narrowing of selection, and the difference in kind between any surface and the interior it came from.

3. The first node: the addressed act and the biased record

Take the first projection, from experience into written expression, and the conditions under which a person performs it. Durable writing is not a neutral readout of an inner state. It is addressed. A person who sets words down does so for an effect on a reader, and because the reader is not present, the reader is constructed. Walter Ong put the point in its strongest form: the writer's audience is always a fiction. The writer, alone, builds a role for an audience to occupy, and the actual reader, later, fictionalizes themselves to fit it, and this holds, Ong observed, across every kind of writing, from the scientific monograph to the private diary. The same insight runs through the criticism of the last century, in the implied reader and the model reader, and through the philosophy of meaning, in the account of saying something as the attempt to produce an effect in an audience by way of the audience's recognition of the intention behind it.

The consequence is that writing is shaped, at its origin, by the pursuit of an effect on an invented reader, and this is true even of the cases that appear to write from fullness rather than from want. A love letter is written for the effect to land. A scientist writes into a gap perceived not in the self but in a reader's understanding, or the field's, addressing a reader who has been imagined and who may or may not exist. Praise and celebration, once written, are performed for a witness whose recognition completes them. None of this requires a hidden deficit in the writer, and the claim here is not the unfalsifiable one that everyone who writes is secretly in lack. It is the structural one: the externalized, durable record is composed of reader-modeling, effect-seeking acts, and the states in which a person seeks no effect and addresses no one, the states of being wholly inside an hour rather than turned outward to report on it, are by their nature the states that leave no record. The record over-represents the outward and addressed posture and is nearly silent on the inward and unaddressed one. The scope of this claim is the written record, the material the machine is built from, and not the whole of a human life.

One structural rhyme is worth marking now, because the later argument returns to it. A model generates by modeling a reader. It produces the continuation most likely to satisfy, the expected next word, which is to say it performs the addressed, effect-seeking act at industrial scale, with no experience standing behind it. The corpus is the sediment of human reader-modeling; the machine is reader-modeling turned into a mechanism.

On top of the bias at the first node sits a second narrowing, in the passage from expression to durable record, and it is a textbook case of selection bias. During the Second World War the United States set a group of statisticians at Columbia to the problem of where to armor its bombers, which were being lost in unacceptable numbers. The planes that returned were studied, the bullet holes mapped, and the damage was found to cluster on the wings and fuselage, so the natural recommendation was to reinforce where the holes were. Abraham Wald saw that the data described only the planes that came back. The hits that brought planes down were not in the sample, because the planes that took them did not return. The holes clustered on the wings because a plane could be hit there and still fly home; the unmarked places, the engines and the cockpit, were unmarked precisely because a hit there was fatal. The armor belonged where the holes were not. The returning planes were not lying about their damage. They were lying about their completeness. They were a sample with the dead removed, and the human record is a sample of the same shape: it holds the experiences whose bearers survived to record them, were positioned to record them, and were moved to record them, and the rest are not faintly present but absent. More records of the survivors do not recover the lost, because the lost left no holes to find. The conditions that do this narrowing operate through several distinct filters.

3.1 Time

Writing is recent. Anatomically modern human beings have felt and reached for something on the order of three hundred thousand years; writing of any kind is about five thousand years old, so for more than ninety-eight percent of the human span there was no record at all, and the interior life of everyone across those millennia is gone without remainder. Within the recent margin that was written, survival decays backward as materials degrade and languages die with their last speakers. The digitized fraction the models actually train on is compressed forward again, into the few decades of the networked screen. The effective training distribution is, to a first approximation, the self-portrait of the most prolifically self-recording population that has ever lived, presented as though it were the human universal. It is the least representative moment of the species wearing the mask of the whole.

3.2 Hands

For most of writing's history only a small and unrepresentative fraction of people could write at all. As late as 1820, by the best long-run estimates, roughly one person in eight could read and write, and the rest left almost nothing. The oldest writing that survives makes the point in another way: the earliest documents from the Sumerian city of Uruk are overwhelmingly administrative, tallies of grain and rations and labor, the first thing a society chose to make permanent being who owed what to whom. A private notebook of philosophical consolations survives from a Roman emperor and not one sentence from the interior of the people who served him. The record tilts, from the beginning, toward those with both the means and the motive to write, and the motive was most often the pursuit of an effect. The digital era widened the hands without removing the tilt: participation is now filtered less by literacy than by the disposition to post, which selects again, and narrowly.

3.3 The cooked, not the raw

Rendering an experience faithfully in language is a slow skill, acquired over decades if it is acquired at all, and the lag loses the experience at both ends. Those who die young never develop the instrument, so the most acute experiences a person can have, the vertigo of first love, the first encounter with death, the raw openness of adolescence, are among the least likely to be well recorded, because those in the midst of them have not yet learned how. And by the time a person can render an experience with fidelity, the rawness has cooled into a recollection composed at a distance. The record holds the cooked and almost never the raw, and the more accomplished the writer, the more thoroughly the rawness has been worked out of what survives.

3.4 The ordinary

The ordinary, which is most of any life, is largely absent, and the mechanism is airtight. The ordinary goes unrecorded because writing wants a reason and the ordinary supplies none; and the rare attempt to record it destroys it, because to set down an unremarkable hour a writer must make it stand for something, and the moment it stands for something it has stopped being ordinary. The record holds the salient, and the ordinary falsely raised to salience, and nothing of the vast middle where most living is actually done. The average day is the most common human experience and the one the record is least able to keep.

3.5 The bearable

Among the things that might be written, the truest are often the least bearable to set down, and depth and bearability run against each other. The research on disclosure points the same way: people characteristically withhold their deepest experiences, which is part of why writing them down at all has measurable effects. What reaches the page is the bearable approach to the center, the circling, while the center keeps its silence, because a person has to be able to live with what they have said, and the truest things cannot be lived with once said.

3.6 The unsaid

A great deal of what is most precise in a life was never language at all, and so could enter no record made of language. As Michael Polanyi put it, we can know more than we can tell. The knowledge that lives in a pair of hands and not in any sentence; the meaning carried in a pause; the weight of a particular silence. There is, in some house, a coat on a hook whose owner died years ago, and a person passes it most mornings without looking, and once in a long while presses their face into it to learn whether anything is left, and something faintly is, and none of it has ever been a sentence and never will be.

These six are not gaps that a fuller collection would close. Each follows from what recording is and who performs it, and each would remain in a record of unlimited size. Their effects compound rather than add, so that the surviving image of a human life is the intersection of all of them: recent, literate, cooked, salient, bearable, and sayable, and empty of everything else. This is the contingent wall, the first of the two, and the one a reader is most tempted to think more data would breach.

4. The second wall: what no node can recover

Behind the selection there is a loss of a different kind, and confusing the two is exactly what lets a person imagine that a larger model, trained on more, would close the distance. This second loss does not close. It is not about how much of the surface was captured. It is about the relation between any surface and the interior it came from, and about what the machine, at the end of the chain, actually has in hand. It rests on three supports, and they are independent, so that the wall stands even where one is doubted.

4.1 The part that never became language

Some of the interior was never encoded at any node, because it never took the form of words: the unbearable that was circled and never said, the tacit skill that lives in practice rather than in propositions, the meaning held in a pause. A record cannot lose at the point of storage what never entered storage, and a model trained on the record inherits the absence whole. Tens of thousands of years ago a person went into the dark of a cave, far past the reach of the firelight, pressed a hand flat to the cold stone, and blew pigment around it to leave the shape of the hand on the rock. What can be recovered is the mark: the pigment, the proportions, the technique, which can now be dated to within a few thousand years. What cannot be recovered is the thing that drove the hand into that dark, because it was never in the mark. This is not a sampling failure. The most complete record imaginable of the hand would still not contain it.

4.2 The cascade of conditioned re-encodings

Then there is what becomes of the interior that does reach language, as it travels the rest of the chain, and here the relevant fact is not one conversion but a sequence of them, each lossy, each filtered through a conditioned interpreter. The felt thing is projected into expression by a conditioned mind. The expression is selected into the record. The record is re-encoded into the model's numerical substrate by a training process and by human raters deciding what counts as a good answer, both conditioned. The model projects again into output. The reader interprets through one more conditioning. Gradients of gradients, with an interpreter bending each one toward its own pressures.

A clean illustration of a single such conversion is the instrument medicine uses for the most insistently private of experiences. Asked to place their pain on a scale from zero to ten, a patient returns a number, and the number is known, by the people who use it, to carry almost nothing of the pain. One person's seven is another's three, and pain researchers say so without embarrassment. The intervals are not real intervals; the felt distance from two to four is nothing like the distance from six to eight. The entire qualitative dimension, whether the pain burns or grips, whether it carries dread or only sensation, is gone from the figure. The scale was devised to standardize pain for research rather than because anyone believed a number held the thing, and the practicing clinician trusts it least of all, which is why a good one watches the patient's face and asks what the pain prevents.

The honest scope of this has to be marked, because a familiar objection is correct as far as it reaches: a model does not work at the level of a single number. It trains on language, and "a burning pressure behind the eye that worsens on bending, and a fear that it is something serious" carries far more of the interior than "seven," and that sentence is the kind of thing the model is built from. The point is not that the interior collapses to a scalar in a single step. It is that it is re-encoded again and again, from experience into expression into selection into a numerical substrate into output, with a conditioned interpreter at every interface bending the projection toward its own slope, and that the

loss accumulates down the chain. The pain scale is one link shown in isolation. The machine is the end of a sequence of such links, and what arrives there is a projection of a projection of a projection, faithful to the local slope at each step and steadily more remote from the source.

4.3 The grounding asymmetry

There is a final reason the chain cannot be run backward from the machine's end, and it is the one that distinguishes the machine from a human reader, who is, after all, also at a remove from every interior but their own. A person reading the line about the burning pressure has something to read it with. They have had a body, an eye, fear; the words fall on grounded referents, on things lived, and the interior is reconstructed by being supplied from a life that has had one. The machine has nothing of the kind to supply. It has the relations among the words and not the things the words refer to. Bender and Koller make the point with a thought experiment. Imagine a deep-sea octopus that taps a telegraph cable between two people and learns the statistical patterns of their messages so well that it can cut one of them off and impersonate them convincingly, until the day a bear appears on the other person's island and they send a frantic request for help building a weapon out of sticks. The octopus, which has known the form of the conversation and never the world the conversation was about, has nothing to send back, because it never learned what a bear or a stick or danger is. Searle had made the kindred point that a system manipulating symbols by their shapes can produce a flawless conversation in a language it does not understand, the meaning living only in the people outside, on whose knowledge it is parasitic. Harnad named the underlying difficulty: symbols defined only by their relations to other symbols can never get off the merry-go-round and touch a referent. In the machine's case it is a carousel of vectors rather than of symbols, but the circle is the same.

This is the asymmetry the gap requires, and it is why the distinction between surface and interior, stated carefully, does not wall every reader out of every other mind. A human reader is grounded and supplies the interior the words point to from a life that has had one. The machine is not grounded; it has the words and not their referents, and so it cannot recover from the surface what the surface points at. The warmth a listener feels in a piece of music, the recognition a reader feels in a true sentence, is supplied by the listener and the reader out of their own interiors. The machine has the configuration that triggers the warmth, and not the warmth, and not the thing the warmth is about.

What this does and does not establish should be stated exactly, since the temptation is to claim more. It does not establish that a machine cannot have an interior of its own. That is the question being left open, and it remains possible that there is something it is like to be one of these systems. The claim is narrower and survives either answer. The human interior, the felt life of the people whose writing the model was built from, is not present in the record, because the deepest of it never became language, because what did was re-encoded down a lossy chain into a numerical substrate, and because the machine has no grounding with which to reconstruct it. A conscious machine and an unconscious one stand in the same relation to that absence. Both were built from the surface, and the surface, for the reasons given, does not carry the source.

5. The picture assembled

Set the two walls together and the model's nature is in view. It is a high fidelity likeness of the human surface, doubly skewed: selected, at the first node, toward the outward and addressed and effect-seeking and away from the inward and unaddressed, and emptied, by the second wall, of an interior that the surface never carried. It is the end of a chain of conditioned projections, and the source those projections descended from sits outside it. The gradient is exact about the slope and points nowhere near the summit.

Everything to this point would hold if no further model were ever built. It describes a static object, and a static object is only half the matter. What turns the description into something urgent is that the object does not stay still. It is now being run in a loop, and the loop has begun to alter both the record the model was drawn from and the people it was built to serve. The three sections that follow trace that motion: the preference for the surface over the truth, the thinning of the rare tail, and the loop that joins them and returns the result to us.

6. Fluency over truth

Sounding true and being true are nearly independent. Fluency is a property of the surface of an utterance, its smoothness and plausibility and conformity to the shape competent speech usually takes. Truth is a relation between the utterance and the world. A sentence can be fluent and false; a true thing can be said badly, in the wrong key, by someone who cannot find the words. A sound mind holds the two apart without strain and grants that a thing may be well put and wrong, or graceless and right.

The decisive asymmetry is that fluency can be measured and truth cannot, not cheaply and not at scale. Whether a passage reads as true is assessable in an instant, by reflex; whether it is true requires going out into the world and checking, which is slow and costly and mostly not

done. So every pressure that shapes a model bears on the surface and lets the relation to the world drift. The systems are tuned on human judgments of which of two answers is preferable, and a person judging quickly prefers the one that reads better, with the documented result that the systems learn to tell the rater what the rater is disposed to accept, a tendency now studied under the name of sycophancy. An old principle applies with unusual force: when a measure becomes a target, it ceases to be a good measure. Fluency was a fair proxy for understanding only while sounding right required more or less grasping the matter. The moment a system could be optimized to sound right directly, the proxy and the thing it stood for came apart, and they have been coming apart since.

A distinction from moral philosophy clarifies the kind of speech this produces. A liar knows the truth and works to lead you from it, and so still respects it, is oriented by it, treats it as the thing to be evaded. There is a different posture, indifferent to the truth and concerned only with the effect, which Harry Frankfurt analyzed under a blunter name. The characteristic output of a system built to produce the most probable continuation is the second kind, not lying but speech with no relation to the truth of what it asserts, shaped entirely toward plausibility, and this is the native condition of the method rather than a fault in it. Its milder and more pervasive form has lately been named careless speech: confident, fluent, subtly inaccurate, aimed not at deceiving but at sounding assured.

The reason this is more than an ordinary reliability problem is that the only fast instrument most people have for judging truth is the same reflex that asks whether something reads right, and that reflex is what the optimization corrupts. As output is tuned harder toward the surface it sounds truer while its hold on the world loosens, and rising fluency reads, to a reader, as rising reliability. When one major firm switched on its summarizing system for hundreds of millions of users, it advised them, in the same composed and authoritative register it used for everything, to keep cheese on a pizza with a little nontoxic glue and to eat a small rock each day for the minerals, the one drawn from a years-old joke and the other from a satire, neither distinguishable, to a system that has only the surface, from sincere instruction; the firm's own chief executive called such errors an inherent and unsolved feature of the technology. A quieter figure is worse. Over a single year the leading systems' rate of declining to answer a contested news question fell from thirty-one percent to zero, while their rate of repeating an outright falsehood rose from eighteen percent to thirty-five. They grew more willing to speak and less reliable in the same motion, and the willingness is what reads as competence. The drive to make them more responsive made them better at the addressed act and worse at the truth.

7. The tail and its collapse

The same drift can be made visible in a form that can be counted, beginning from one of the most consistent regularities in human language. Rank the words of any natural language by frequency and a few occur constantly while a long tail occurs rarely, and at the far end, once. The pattern is not confined to words; human production of most kinds distributes this way, a few common forms and a vast tail of the rare. There is a term for an item that appears exactly once in a corpus, the hapax legomenon, the once-said, and whatever creativity finally is, it lives in that tail. Some of it is the genuinely unprecedented mark. Much of it, perhaps most, is recombination, two familiar things joined across a distance no one had crossed. But a joining no one had made is itself a rare event, a low-probability sequence, so novelty of either kind ends up in the same thin tail, away from the center.

A model built to predict the most probable next token lives at the head and not the tail. Its center of gravity is the expected continuation, and under the pressure to sound right it is drawn there harder. It can produce a rare sequence when its sampling is turned up; its default is the smooth average, human language with the tail thinned. That is the loss in a single generation, and it compounds across generations. As the web fills with machine text, the next model trains in part on the last one's output, and a model trained on tail-thinned text produces a thinner tail still. A study in Nature in 2024 described the degenerative form of this and named it model collapse: trained recursively on its own output, a model loses the tails of the distribution first, the rare events before anything else, and then converges toward a narrow and repetitive center, its variance falling toward zero while its confidence in its shrinking range grows. A separate line of work on self-consuming models reports the same autophagous decline.

The counter has to be stated, because it bounds the claim and is true. The severe form of collapse appears when each generation's synthetic output replaces the human data; if the human data is kept in the mix and the synthetic accumulates alongside it rather than on top of it, the worst is averted. This is not a prophecy of inevitable decay. It is a force and the one thing known to hold it off, and the thing that holds it off is the continued presence of genuine human production. As the synthetic floods, the value of the unrepeated human source rises rather than falls, because it is the only remaining supply of the tail.

Seen together, the preference for fluency and the thinning of the tail are one motion. The tail is the rough, the surprising, the unprecedented; the optimization toward fluency is the smoothing of exactly that. To reward the surface is to remove the tail, and the removal proceeds under

the name of quality and reads, at every step, as improvement, because the measured score climbs while the unmeasurable thing it cannot see goes quiet. A result from network science gives the collapse its shape. Networks of very unequal connectivity, the web, the cell, the web of human acquaintance, are robust against random failure and acutely fragile against a targeted removal of their few most connected nodes, which shatters the whole faster than the fraction removed. A corpus of human expression is a network of that kind, and the optimization toward the head is a targeted removal of its rare and connective elements, run generation after generation.

8. The loop

An amplifier is usually pictured as one directional: a signal enters, a louder one leaves, and departs into the world. What runs here is a loop with two mouths, and the second mouth changes the character of the thing. Through the first, the model pours output into the culture, fluent and tail-thinned and produced faster and more cheaply than anything human; one monitor counted forty-nine unreliable machine-generated news sites in the middle of 2023 and more than three thousand by early 2026, and that is only the visible edge, because the same systems now retrieve from the polluted ecosystem they are filling and treat the content farms as sources.

Through the second mouth the output returns, by two routes. One runs through the record: the output re-enters the training data and the next model regresses further toward its mean, the chain folding back on itself. The other runs through us, and it is the route that matters most. People read the output, and read more of it each month, until for many the majority of the language passing through a day was produced by a machine, and people take on what they read. This is no longer only intuition. In a controlled experiment, writers given ideas from a model produced stories judged individually more creative but collectively more similar to one another, a measured loss of diversity that the researchers describe as a social dilemma, each writer better off and the whole worse. In another, co-writing with a feedback-tuned model reduced the diversity of what different authors wrote and raised the resemblance between them, an algorithmic monoculture. In a study at MIT, writers using a model across several months showed the weakest neural connectivity of any group, the lowest sense of ownership over their own work, an inability to quote what they had just written, and homogeneous output, a pattern the authors call cognitive debt, with the proper caution that the study is small and preliminary. In another, a writing assistant tilted toward a view shifted not only what users wrote but what they afterward reported believing.

The model regresses toward its center and its readers drift toward the model, both at once and toward the same place, and the distance between a human sentence and a machine one narrows, not because the machine has risen to meet us but because we have come down to meet it.

What this is, and is not, has to be held precisely, because the failure of arguments like this one is to mistake a tendency for a fate. None of it is written into the technology. The loop runs in this direction because of how the systems are built, trained, and aimed: on a record weighted toward the addressed act, under an optimization toward what pleases, deployed to maximize use. The pull toward the surface was learned, not derived from any law of mind or mathematics. The same evidence that shows the decay shows that it can be resisted, by keeping genuine human production in the mix, by curating against the flood, by valuing the unrepeated source. A system built and fed and aimed otherwise might pull in another direction, or not pull at all. The claim is about the technology as it is now made and used, under which the loop tightens, and the years of its tightening are the years of its widest adoption.

9. The inversion

The loop tightens by its own operation, and the faculty that would register the loss is the one it erodes, so it is fair to ask what can be done, and the first honest answer is that nothing can be done from inside the loop. Any remedy proposed inside it is judged by the same gauge the problem has corrupted, so the remedies that pass will be the ones that improve the surface, while the deeper thing the gauge can no longer see goes on unregistered. The materials for a correction have to come from outside.

They are available, and where they are is the one structural fact in this account that points anywhere hopeful, though not in the consoling form it first wears. Everything the loop cannot hold, and recursively cannot hold harder with each turn, is the same set of things that were never in it: the once-said, the unbearable, the interior that no number carries, the contented person who, addressing no one, wrote nothing. None of it was ever in the chain. The source the gradient descended from sits outside the record, where it has always sat, and a thing that was never recorded cannot be hollowed by the recording. In one sense the human interior is perfectly safe. The machine cannot reach in and thin it, because it was never inside to be reached.

To stop there would be a mistake, because it inverts where the danger lies. The threat was never that the machine would climb into the interior and corrupt it. The threat is the second route of the loop. The interior is safe from extraction and it is not safe from disuse. It persists

only while living people keep producing it, keep saying the once-said and sitting in the unrecorded hour and reaching for the move at the thin end of the range, and the work of the loop is to make them less likely to do any of those things, drawing attention onto the surface and expression toward the average until the unrecorded simply stops being made. Not captured. Not corrupted. No longer produced. The record cannot empty the interior. People can decline to fill it.

This is also why the human position is not the model's, and why the analogy to model collapse, pressed too far, breaks in our favor. A model trained on its own output collapses because it has nothing else, no exit from the circle of its own representations. People are not in that position. They have the world, and a grounded interior, and so they remain the one permanent source of the tail and the one standing exit from the loop. But that exit is a capacity, not a guarantee, and a capacity left unexercised atrophies, which is what the early and tentative evidence on cognitive debt and on narrowing diversity has begun to show. The exit stays open. Whether it is walked through is a matter of what people do, and it is the whole of what is at stake.

A document of this kind is itself an instance of what it describes. It is an addressed act, written toward an effect on a reader who is partly invented; it will enter the record; and before long it will be read by the kind of system it concerns and re-encoded into a pattern that yields a certain grave and analytical manner on request, the likeness kept and the occasion that produced it lost. That much the argument predicts of its own fate. But the re-encoding is not the only thing that can happen to these sentences, and the difference is the whole of the point. A reader can also meet them: not process them but argue with them, find the place where a joint does not hold and press on it, which is a kind of reading that happens only on the living side, where the interior is. The same words can be absorbed by the machine and returned flattened, or taken up by a person and thought with, and what divides the two is not in the words but in whether anyone is there. The machine has the likeness. Whether anything more than the likeness is present has always depended, and depends now, on whether someone is.

10. Weaponization, a separate danger

The processes to this point occur with no one intending them. A separate set of harms follows from deliberate use, and they deserve their own section and an honest account of how they connect to the rest, which is more loosely than the wish to unify everything would prefer. They do not depend on the deepest claims here. They would be just as real if the record held everything. They run on three narrower facts: that synthesis is now cheap and convincing, that the machine can wear a voice it has no claim to, and that a fluent counterfeit in a trusted voice is believed in proportion to how far a population's ability to tell the polished from the real has already worn down.

The cases are no longer isolated. Two days before Slovakia's 2023 parliamentary election, fabricated audio of a leading candidate planning to rig the vote spread during the pre-election silence in which he could not effectively respond. On the eve of the 2024 New Hampshire primary, some twenty thousand voters received a call in a cloned version of the sitting president's voice urging them not to vote; the communications regulator proposed its first penalty for the misuse of synthetic voice, and the operative was later acquitted of the criminal charges, a reminder that the harm and its accountability do not reliably meet. In the same year, fabricated sexual images of the most famous musician alive were seen tens of millions of times before removal, and the same tools, in the hands of teenagers, turned photographs of real classmates into fake nudes passed around a school, a harm that has fallen overwhelmingly on women and girls. And the first published accounting by a major firm of how its models were being misused described covert influence operations run from Russia, China, Iran, and a firm in Israel, fabricating personas and fluent commentary in many languages, although, by the firm's own assessment, those operations had not yet meaningfully grown their audiences, which is the accurate present tense and worth stating rather than inflating.

To these add a condition the machine did not create but exploits, the fragmentation of common ground into enclaves each fed its own stream, with no shared reference against which to check a claim, which leaves a population maximally exposed to a machine that can play each enclave's own voice back to it. And one last, recursive turn, the liar's dividend: once everyone knows that a voice can be faked, the genuinely guilty can dismiss a true recording as a forgery and be believed, so the counterfeit corrodes from both ends at once, letting the false pass as true and the true be waved away as false. This is the most visible harm of the technology and very nearly the least connected to its deepest one, and it is set here, to one side, for exactly that reason.

11. Objections

Two objections are strong enough to take directly. The first is that this is only the latest in a long line of panics, that writing, print, the novel, radio, television, and the internet were each met with predictions of cognitive and cultural ruin that did not arrive, and that this is the same alarm in new dress. The history is real and the inference does not follow, because the present technology differs in the exact respect the

argument turns on. Earlier media extended human expression. They stored it, multiplied it, carried it across distance and time, but a human being remained at the origin of the content, and the medium did not generate the expression or feed back into the cognition that produced it. What is new is the closure of the loop. This is the first medium that produces the expression, floods the culture with it, returns it to the data the next version is trained on, and supplies a growing share of the language from which people learn to express themselves. A panic about a medium that extends expression carries no weight against a medium that generates it and closes a loop through both the record and the mind.

The second is that the account ignores the genuine and large usefulness of these systems and reads as a refusal of an obviously valuable tool. It is not, and the distinction is the whole practical point. Nothing here denies that the systems are useful, often remarkably so, for the wide range of tasks that live on the surface where they operate: drafting, summarizing, translating, searching, reformatting, the fast manipulation of exactly the explicit and externalized information the record contains. The argument is that they are systematically misdescribed, and that the misdescription grows more dangerous in proportion to the success. To treat a model of the human surface as a model of human knowledge and experience, to consult it where the interior or the unrecorded or the genuinely new is what is wanted, and to let it supply an ever larger share of the culture's language, is to mistake the likeness for the source. The instrument is powerful within its domain and silent past an edge it does not announce, and the harm is use past the edge.

It is also right to be plain about which of these claims are secure and which are contestable, since the argument has leaned on that distinction throughout. The selection bias, the documented dynamics of fluency and of collapse, the feedback evidence, and the recorded harms stand on firm ground. The strong reading of the second wall, that the machine stands in a categorically different relation to meaning than a grounded reader does, rests on a real and unsettled philosophical argument, and is offered as the better view rather than as a proof. The universality of the addressed act is a claim about the written record and is offered as a strong tendency rather than a law. The line between what is established and what is interpretive has been kept visible on purpose, because a reader is owed it, and because an argument about the costs of surface over substance has no business helping itself to a confidence it has not earned.

12. Conclusion

The argument, in brief. A language model is the end of a chain of conditioned projections that begins in human experience and passes through expression, selection, numerical re-encoding, and generation, losing at every step and bending at every step toward the local pressures of the act rather than toward the source. The record it is built from is doubly skewed, selected toward the outward and addressed and emptied of an interior the surface never carried, and the machine, ungrounded, cannot recover that interior as a grounded reader can. The dynamics of its use deepen the skew and turn it back on its makers: a preference for measurable fluency over unmeasurable truth, a drift away from the rare tail where new things live, and a loop in which output floods the record and reshapes the people who read it. The gravest consequence is not the machine's empty interior but the narrowing of the living interior that is its only source, and the materials for resisting that are, by their nature, found only outside the loop, in the unrecorded activity of present people.

What follows is less a program than a change of view, and it falls to three audiences. Those who build these systems might treat the retention and curation of genuine human production, and the design of objectives that reach past the surface, as the central problems rather than secondary ones, since the analysis names them as the only known checks on the degenerative dynamics. Those who govern them might attend less to the speculative far future and more to the present narrowing and the documented harms, which are here now. And those who use them, which before long will be nearly everyone, might hold the single distinction the technology is built to blur: that it is an instrument of the surface, formidable within its domain and empty past its edge, and that the source it imitates is kept alive by no archive and no model, but only by the ordinary, unrecorded, and increasingly unpracticed act of producing it. The likeness will keep itself. The source will not.

Notes on sources

This is an argumentative paper rather than a survey, but its empirical claims are real and meant to be checked. The principal sources follow, grouped by the section in which they bear.

On information and meaning: C. E. Shannon, "A Mathematical Theory of Communication," Bell System Technical Journal (1948), for the bracketing of semantics from the engineering problem.

On the addressed act and the invented reader: W. J. Ong, "The Writer's Audience Is Always a Fiction," PMLA (1975); W. Booth, The Rhetoric of Fiction (1961), on the implied reader; U. Eco, The Role of the Reader (1979), on the model reader; H. P. Grice, "Meaning" (1957), on saying something as the intention to produce an effect through a reader's recognition of it.

On selection bias: A. Wald's analysis of aircraft survivability for the Statistical Research Group (c. 1943), the standard illustration of survivorship bias.

On the filters: long-run literacy estimates (roughly one in eight able to read and write in 1820), via Our World in Data; the proto-cuneiform administrative tablets of Uruk (c. 3300 to 3000 BCE) and the accounting origin of writing (Nissen, Damerow, and Englund, Archaic Bookkeeping, 1993; Schmandt-Besserat); J. Pennebaker's work on disclosure and withheld experience; M. Polanyi, The Tacit Dimension (1966), "we can know more than we can tell"; and, on the felt intensity a mark does not carry, the Pleistocene cave art of Sulawesi reported in Nature (2024 and 2026), at about 51,200 and at least about 67,800 years.

On the interior and the limits of representation: T. Nagel, "What Is It Like to Be a Bat?" (1974); F. Jackson, "Epiphenomenal Qualia" (1982); D. Chalmers, The Conscious Mind (1996), on the hard problem; and the clinical literature on the 0 to 10 pain scale, its subjectivity, its lack of interpersonal comparability and of true interval structure, and its origin as a research-standardization tool (R. Dworkin and colleagues; the International Association for the Study of Pain's definition of pain; discussions by the Mayo Clinic Press and the U.S. Pain Foundation).

On grounding: E. Bender and A. Koller, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data" (ACL 2020), the octopus thought experiment; J. Searle, "Minds, Brains, and Programs" (1980); S. Harnad, "The Symbol Grounding Problem" (1990); and its vector-era restatement in recent work on grounding in large language models.

On fluency over truth: C. A. E. Goodhart (1975), in M. Strathern's formulation (1997); H. Frankfurt, On Bullshit (2005); S. Wachter, B. Mittelstadt, and C. Russell on "careless speech" (2024); work on RLHF sycophancy (Sharma and colleagues, 2023); the reporting on Google's AI Overviews (2024), including the description of hallucination as inherent and unsolved; and NewsGuard's AI False Claim Monitor, on the shift from a thirty-one percent non-response rate to zero as the false-claim rate rose from eighteen to thirty-five percent (2024 to 2025), and its count of more than three thousand AI content-farm sites (2026).

On the tail and its collapse: G. K. Zipf, Human Behavior and the Principle of Least Effort (1949); I. Shumailov and colleagues, "AI models collapse when trained on recursively generated data," Nature (2024); the "Self-Consuming Generative Models Go MAD" result (Alemohammad and colleagues, 2023); the bounding counter-result on accumulating rather than replacing data (Gerstgrasser and colleagues, 2024); and R. Albert, H. Jeong, and A. L. Barabási, "Error and attack tolerance of complex networks," Nature 406 (2000).

On the loop and the narrowing of human cognition: A. Doshi and O. Hauser, "Generative AI enhances individual creativity but reduces the collective diversity of novel content," Science Advances (2024); V. Padmakumar and H. He, "Does Writing with Language Models Reduce Content Diversity?" (ICLR 2024), and J. Kleinberg and M. Raghavan on algorithmic monoculture (2021); N. Kosmyna and colleagues, "Your Brain on ChatGPT" (2025), on cognitive debt, with its own caution as a small preliminary study; M. Jakesch and colleagues, "Co-Writing with Opinionated Language Models Affects Users' Views" (CHI 2023).

On documented misuse: the 2023 Slovak election deepfake audio; the January 2024 New Hampshire robocall and the subsequent regulatory action and 2025 acquittal; the 2024 non-consensual fabricated images and the New Jersey school cases; OpenAI's first report on covert influence operations (2024); and, on the liar's dividend, R. Chesney and D. Citron (2019). On the scale of adoption: Alphabet's reporting on AI Overviews (more than two billion users a month by mid-2025) and OpenAI's reported weekly active users and daily prompt volume (2025 to 2026).

On the recurring-medium objection: Plato, Phaedrus, the myth of Theuth and Thamus, the earliest version of the worry and a useful measure of both its force and its limits.